Neural Network Seminar
Multi-modal Speech Representation Learning

Carlos Schmidt
uuwss@student.kit.edu

Abstract

While speech representation learning is a well-known approach to train versatile models capable of solving different natural language processing (NLP) tasks like automatic or visual speech recognition (ASR/VSR), a relatively new path towards leveraging multiple modalities shows promising downstream performance gains. Furthermore, multi-modal data like video with sound is everywhere and because of their intuitiveness easy to produce, e.g., by recording a video of a person speaking. The reviewed models in this paper exploit different combinations of three types of modalities: audio, video and textual data. The performance gain of these models with respect to previous single-mode approaches is shown as well as differences between the models themselves. Overall, multi-modal approaches to speech representation learning demonstrate enormous potential and have already set new benchmarks for numerous NLP tasks.

1 Introduction

The goal of learning intermediate latent representations of a given input instead of learning an explicit task end-to-end is to be able to extract information out of the input and use this information to tackle multiple tasks. Speech representations are latent representations of speech input. This input could be in formats such as raw waveform, or mel filterbank features. Speech representations are related to (textual) semantic word embeddings as these embeddings are also latent representations with extracted information, in that case their semantic relatedness (Pennington et al., 2014). The difference from textual embeddings to speech representations and what makes learning speech representations hard, is that spoken words don't have clear boundaries, speech input is continuous as opposed to discrete textual words and that speech contains more information than text like for example speaker information, noise or emotion. The first attempts at learning speech representations were done using clustering algorithms like k-means or Gaussian Mixture Models (Rabiner and Wilpon, 1979) and later improved by adding Hidden Markov Models to allow processing of continuous speech rather than single words (Bahl et al., 1986). Currently, the prevalent approach for learning speech representations is to perform pretext task optimization, also known as pre-training. In this approach, representations are learned by solving a task that is derived by the structure of the underlying unlabeled data like for example predicting the next word in a sequence of words or predicting a masked word in a sequence. The advantage of this approach is that it only needs unlabeled data which is of a higher availability than labeled data. Within the pretext task optimization approach, the following three learning paradigms can be considered the most widely used:

1. *Contrastive learning*: Contrastive learning is based on the idea of understanding the underlying structure and relationships within the data by emphasizing the differences and similarities between different examples of it. This means, a model needs to find representations in the latent space, such that semantically similar samples of the training data have a high similarity in the representation space while semantically dissimilar samples have a low similarity in the representation space. A common training objective is the SimCLR loss (Chen et al., 2020):

   $$\mathcal{L}_{\text{SimCLR}}^{(i,j)} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

   Here, z_i and z_j are similar samples, while z_i and z_k are dissimilar. $sim(\cdot, \cdot)$ is the cosine similarity.

2. *Generative learning*: In generative learning, an input is reconstructed based on a limited view of it. This can mean predicting the next

word in a sequence, a masked word within an input or predicting the original from a noisy input, as can be seen in (Vincent et al., 2008). A typical training objective is the basic reconstruction, or L_1-Loss.

3. *Predictive learning*: In predictive learning, typically another model is used to compute pseudo labels for the model to train with. An example of this, which will be discussed later in more detail, is the “Hidden-Unit BERT” (HuBERT) model, which initially computes its training targets with the k-means algorithm and later uses intermediate representations of its inner structure as labels.

This paper is structured as follows. First, single-mode speech representation models are introduced, specifically wav2wec2.0 and HuBERT, which are the foundation and basis of some of the multi-modal speech representation models. Afterwards, multi-modal speech representation models are shown by first looking at their architecture, then the learning approach and some of its specifics and finally the performance of the models is discussed with respect to other multi-modal and also single-mode models. Finally, a discussion about the impact, advantages and challenges of multi-modal speech representation learning is conducted including a comparison to single mode approaches.

2 Single-mode Speech Representation Models

In this section, we briefly discuss two different approaches for single-mode speech representation models. Both approaches use audio data to learn representations and can outperform previous supervised models in tasks such as automatic speech recognition using the same amount of training data (Baevski et al., 2020). A more comprehensive list of single-mode approaches to speech representations can be found in Mohamed et al. 2022.

2.1 wav2vec2.0

The wav2vec2.0 model by Facebook AI (Baevski et al., 2020) consists of a CNN + Transformer architecture. The CNN takes as input raw audio data and outputs latent speech representations. These representations are on the one hand quantized and used as pseudo labels for the training procedure. On the other hand they are given as input to the Transformer network, which in turn outputs contextualized representations. Then, a contrastive loss is applied to different parts of the Transformer’s outputs, after masking said outputs. This means, a combination of contrastive, generative and predictive learning is applied. The goal is to maximize the similarity between the contextualized representations and the localized representations.

2.2 HuBERT

The “Hidden-Unit BERT” (HuBERT) model also consists of a CNN followed by a Transformer, specifically the BERT architecture (Devlin et al., 2019). The difference to the wav2vec2.0 approach is that HuBERT uses in its first iteration k-means clusters, and in further iterations clustered intermediate layer features of its Transformer network as its learning targets. Furthermore, HuBERT does not employ a contrastive loss but rather uses the predictive and generative learning paradigms by predicting the clustered targets at masked positions of the model’s output (Hsu et al., 2021).

3 Multi-modal Speech Representation Models

This section presents three examples of speech representation learning models, each leveraging a different combination of input channels. First, we will look at SpeechT5 (Ao et al., 2022), which uses additionally to the audio modality the text modality for learning speech representations. Next, we will take a look at AV-HuBERT (Shi et al., 2022), which is an extension of the single-mode model HuBERT (Hsu et al., 2021) and benefits from the audio modality as well as the video modality. Lastly, VAT-LM (Zhu et al., 2023) will be covered, leveraging all three mentioned modalities: audio, video and text.

3.1 SpeechT5

The SpeechT5 framework, first introduced in 2021 by Microsoft (Ao et al., 2022), is an expansion of the Text-to-Text Transfer Transformer (T5) framework (Raffel et al., 2023). SpeechT5 is, as the name suggests, an audio and text approach to speech representations.

3.1.1 Model Architecture

SpeechT5 is based on the encoder/decoder Transformer architecture also used in the original Transformer paper (Vaswani et al., 2023).

The encoder component takes as input the audio and text data which are previously encoded with pre-nets. For audio data, a pretrained wav2vec2.0 model is used as pre-net, while for text, word embeddings are used in the preprocessing step. In the actual Transformer Encoder, the preprocessed audio and text are input to produce representations for the decoder.

The decoder takes as input the representations produced by the encoder and outputs a sequence, which is then fed into the post-processing mode-specific networks. For audio data, this is a log Mel-filterbank predictor, and for the text output a representation-to-token transformation network, which computes a probability distribution of tokens from the decoder's output. Finally, for some tasks like voice conversion, the input data is also fed into a decoder pre-net, which is a feed forward network with the log Mel-filterbank features of the input for audio and another embedding layer for text inputs.

3.1.2 Training

There are three types of pre-training for SpeechT5: Speech pre-training, text pre-training and joint pre-training. One of the goals of pre-training is "to learn representations capturing the modality-invariant information" (Ao et al., 2022) for tasks such as ASR. The speech pre-training uses the generative tasks of bidirectional masked prediction and sequence-to-sequence generation with randomly masked portions of the input. The loss functions used are the cross-entropy loss for the masked prediction and the L1-loss for the generation task. For text pre-training, masked prediction is performed as well, where 30% of the text is randomly masked. As a loss function, the maximum likelihood loss is used. Finally, the joint pre-training approach aims to make the audio and text representations (the output of the Transformer Encoder) capture "modality invariant information" (Ao et al., 2022). This is achieved by using a shared codebook for both modalities. The Transformer Encoder representations are converted into codebook elements by finding the closest element, which is done with a nearest neighbor search over the L2 distance:

$$c_i = \arg \min_{j \in [K]} ||u_i - c_j||_2$$

Here, u_i is the Transformer Encoder representation, c_i is the target codebook entry and K is the amount of codes in the codebook. Lastly, a diversity loss is used to encourage the encoder to use more codes by smoothing the distribution of the probabilities of each code being used.

For the fine-tuning process, different combinations of pre- and post-nets are activated. For example, when fine-tuning for ASR, the speech-encoder pre-net and the text-decoder pre- and post-nets are used as well as the Transformer Encoder and Decoder. Then, the loss of the specific task is used to train the SpeechT5 model end-to-end.

3.1.3 Performance

In Table 1, a performance comparison between SpeechT5 and the single-mode approaches shown in Section 2 is presented. The results show that SpeechT5 consistently achieves a lower word error rate (WER) than the previously covered single-mode approaches. While SpeechT5 does outperform those single-mode models, it is important to note that SpeechT5 uses an additional 400 million text sentences in the pre-training phase (Ao et al., 2022), which wav2vec2.0 and HuBERT cannot.

Model	LM	test-clean	test-other
HuBERT	-	5.8	13.3
wav2vec2.0	-	6.1	13.3
SpeechT5	-	**4.4**	**10.4**
HuBERT	4-gram	3.4	8.1
wav2vec2.0	Transf.	2.6	6.3
SpeechT5	Transf.	**2.4**	**5.8**

Table 1: ASR Performance on the 100 Hours Subset of LibriSpeech. (from Ao et al. 2022)

3.2 AV-HuBERT

Audio-Visual Hidden Unit BERT (AV-HuBERT) by Meta AI (Shi et al., 2022) was first introduced in 2022 and combines audio as well as video input to learn speech representations. As the name suggests, AV-HuBERT is based on the concepts of HuBERT (cf. Section 2.2), also using clustering for target generation and predicting these targets from masked inputs.

3.2.1 Model Architecture

In addition to the base architecture of the HuBERT model, the authors chose a ResNet-18 CNN as a visual encoder and a linear projection layer for the audio input. To combine the inputs of both the audio and video channels, an "audio-visual fusion" module was added. This module simply concatenates the encoded inputs. After the fusion module, a Transformer network creates contextualized representations of the sequential input.

3.2.2 Training

The training procedure of AV-HuBERT is similar to HuBERT's training procedure (cf. Section 2.2, (Hsu et al., 2021)), combining generative and predictive learning by using masked prediction and learned targets. There were however some tweaks introduced for training AV-HuBERT, which are explained below.

Clustering: The authors tested different inputs for the clustering to produce learning targets: Audio only, video only and a combination of both. The conclusion to which they came was that producing targets with the combination of both channels yields the best performance. Like in the HuBERT approach, the first targets are generated with the k-means algorithm on the Mel Frequency Cepstral Coefficients (MFCC) features.

Modality dropout: Since there are two input channels, a different strategy for the dropout regularization technique was chosen. With probability p_m, neither channel is masked. With probability $(1 - p_m)p_a$, only the video channel is masked and with probability $(1 - p_m)(1 - p_a)$, both channels are masked. We will later see that this technique helps AV-HuBERT in achieving higher scores in visual-only tasks.

Masking strategy: Besides the modality dropout, the inputs of both channels are randomly masked before the channel-specific encoders. Instead of masking the visual inputs with white noise or zeroing the frame segments however, the authors proposed to replace them with other segments from the same input stream. The impact of this method is also shown later in the performance section.

3.2.3 Performance

The primary results for AV-HuBERT is the performance for visual speech recognition (VSR), or "lip-reading" and can be seen in Table 2. In this task, AV-HuBERT achieved a big improvement compared to the state-of-the-art (SOTA) approach by Makino et al. 2021. Furthermore, the previous SOTA model was trained in a purely supervised fashion with 31,000 hours of labeled data from the "Lip Reading Sentences 3" dataset (Afouras et al., 2018), while AV-HuBERT only used 433 hours of labeled data and an additional 1,759 hours of unlabeled data. Also, AV-HuBERT combined with self-training achieved the new SOTA. For the self-training, they used a fine-tuned HuBERT model to generate pseudo-labels for unlabeled videos. These pseudo-labeled videos were used in combination with the labeled data to fine-tune AV-HuBERT.

Model	Type	Un-/labeled (hrs)	WER(%)
Ma et al. 2021	Conformer	-/433+157	43.3
Makino et al. 2019	RNN	-/31,000	33.6
AV-HuBERT	Transformer	1,759/433	28.6
AV-HuBERT + ST	Transformer	1,759/433	**26.9**

Table 2: Visual Speech Recognition (VSR, "Lip-Reading") Performance (Excerpt from (Shi et al., 2022)). Conformer: Convolution-Augmented Transformer, ST: Self-Training.

In ASR, baseline AV-HuBERT cannot achieve better results than HuBERT. However, using the audio-visual clustering targets when pre-training audio-HuBERT (AV-HuBERT with $p_m = 0, p_a = 1$) results in a slightly better ASR performance with a 1.3% WER when compared to a 1.5% WER for the normal HuBERT model (Shi et al., 2022). Finally, the impact of the pre-training techniques on the VSR task is shown in Table 3. For the masking strategy, the authors showed that several other masking techniques also performed worse than their suggested technique.

Technique	WER (%)	ΔWER (%)
Modality dropout	Used 46.8/55.3	Not Used +8.4/+1.7
Masking strategy	Proposed 46.8/55.3	Gaussian Noise +5.6/+2.6
Masking Probability	$a: 0.8, v: 0.3$ 46.8/55.3	$a: 0.8, v: 0.8$ +12.5/+6.3

Table 3: Ablation Results for the Different Pre-training Techniques. The Different Columns Show the Validation and Test Set WER Difference of a Pre-trained Av-Hubert for the VSR Task When Using or Omitting the Technique. (Shi et al., 2022)

3.3 VAT-LM

The Visual-Audio-Text Language Model (VAT-LM), is an approach of speech representation learning that merges three modalities to learn representations. The model was first introduced by Microsoft in 2022 (Zhu et al., 2023). It has structural similarities to AV-HuBERT (see Section 3.2), where the common idea is that the different modalities' representations should not share a latent space (as we have seen in 3.1), rather the fusion of them should be mapped onto a combined representation.

3.3.1 Model Architecture

As previously stated, VAT-LM has a similar architecture to AV-HuBERT, also comprising of a fusion module followed by a Transformer Encoder. The difference here is in the third input channel, the textual modality. Like AV-HuBERT, VAT-LM uses a ResNet-18 for the visual modality and the log-filterbank features for the audio data. The two channels are combined such that four video frames are matched with one audio frame, since the modalities have different sampling rates. Lastly, the textual data is encoded using an embedding layer, where again, the input text is first converted into phonemes to tackle the problem of different sampling rates / frame sizes of the modalities. The fusion module, like for AV-HuBERT, is a simple concatenation operation with the channel's encoded inputs.

3.3.2 Training

Data coming from three modalities can be of several forms such as audio-video, audio-text, audio-only or text-only data. For each of these combinations the authors proposed to create the pre-training targets differently:
Audio-video data: For this combination, a pre-trained AV-HuBERT is used to create hidden units which are then clustered with the k-means algorithm to get the learning targets.
Audio-text and audio-only data: Here, the authors also used AV-HuBERT while omitting the text data for the audio-text input while computing the targets. The visual input needed for AV-HuBERT is set to zero.
Text-only data: Finally, for text-only data, the authors proposed to use transcribed ASR data to train text encodings. Specifically, the transcription of the ASR data is converted to phonemes and then transformed with a phoneme2unit model, while the audio data is converted to hidden units with AV-HuBERT which in turn function as targets for training the text encoder.
After generating the pre-training targets, the authors proposed the "unified masked prediction" pre-training task (Zhu et al., 2023). Here, similarly to HuBERT, the input features coming from the fusion module are masked. After computing the contextual representations with the Transformer Encoder, the targets specific to the modality of the masked input segments are used for a prediction task.

3.3.3 Performance

VAT-LM is compared to different approaches mainly by its audio-visual speech recognition and visual speech recognition performance. Results of this comparison with the previous state of the art, AV-HuBERT, are shown in Table 4. It can be observed, that VAT-LM achieves a lower WER than AV-HuBERT with a similar amount of parameters (approximately 2.1% higher for the large models) (Zhu et al., 2023). However, it must be noted that VAT-LM was able to leverage an additional 600M text data to the audio-visual data, which AV-HuBERT used, for pre-training.

Model	Type	AVSR	VSR
AV-HuBERT	Transformer	1.4	28.6
VAT-LM	Transformer	**1.2**	**28.4**
+ Self-training:			
AV-HuBERT	Transformer	-	26.9
VAT-LM	Transformer	**1.2**	**26.2**

Table 4: AVSR and VSR Performance (WER in %) of VAT-LM in Comparison to AV-HuBERT (Excerpt from Zhu et al. 2023). For More VSR Results, See Table 3. For the Self-Training Explanation, See Section 3.2.2.

4 Discussion

By leveraging a bigger set of modalities than previous, single-mode, speech representation learning approaches, the presented multi-modal models are able to achieve higher accuracies and lower word error rates for several tasks like visual speech recognition, which we showed in this review. A reason for this is that even for single-mode tasks like automatic speech recognition, the additional channels used for the pre-training of these models help enriching the speech representations and make them more robust, because they all share the common concept of the language and speech. Another advantage of multi-mode models is that they often require less labeled data than single-mode approaches as we have seen in Section 3.2. Nonetheless, the performance comparisons may not all be perfectly representative since technically more data – or rather data from more modalities – was used to train the multi-modal models. Other challenges in the field of multi-modality speech representation learning include the available data typically being scarce and domain-specific, e.g., for the visual speech recognition task where a specific pose of the speaker is needed in the video frames for a network to detect the lip movements.

References

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. ArXiv:2110.07205 [cs, eess].

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

L. Bahl, P. Brown, P. de Souza, and R. Mercer. 1986. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 49–52.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers.

Kohei Makino, Makoto Miwa, and Yutaka Sasaki. 2021. A neural edge-editing approach for document-level relation graph extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2653–2662, Online. Association for Computational Linguistics.

Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-Supervised Speech Representation Learning: A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210. ArXiv:2205.10643 [cs, eess].

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

L. Rabiner and J. Wilpon. 1979. Considerations in applying clustering techniques to speaker independent word recognition. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 578–581.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. ArXiv:1910.10683 [cs, stat].

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction. ArXiv:2201.02184 [cs, eess].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Pascal Vincent, Hugo Larochelle, Y. Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096–1103.

Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2023. VATLM: Visual-Audio-Text Pre-Training with Unified Masked Prediction for Speech Representation Learning. *IEEE Transactions on Multimedia*. ArXiv:2211.11275 [cs, eess].